From Philosophy to XAI (via Psychology)

Mark T. Keane

University College Dublin, Dublin, Ireland

Counterfactual Perspectives

- ***Philosophy*** :
  - How do we handle them in a Truth-Functional way?
  - What Logic/Logics do we need?
  - Role in Scientific Causality?
- ***Psychology***:
  - Thinking about the past, future, and causes (with distinct associated emotions)
  - What gets changed? Mutability, fault lines of reality
  - Theories of counterfactual thinking…

PHILOSOPHY

The Best of All Possible Worlds…

Logical Issues: Truth

*"**If** kangaroos had no tails, **then** they would topple over"*

- Counterfactuals can be true or false but establishing that truth is non-trivial (Nickerson, 2015)
- Quine says that they may not be truth-functional
- **Problem**: The ***if-part*** is false and the ***then-part*** is false, so truth cannot be function of its components (and yet?)
- **Solution**: Counterfactual is true in a *possible world,* that is otherwise the same as this world (Stalnaker,1968)

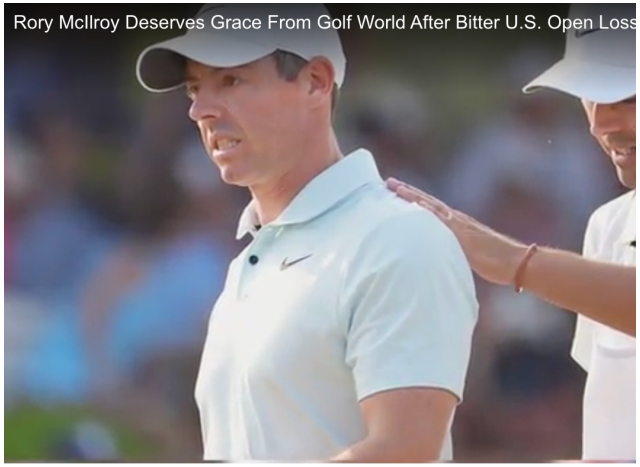
Byrne (2016) Annu. Rev. Psychol.

Counterfactual is the Closest Possible World, Minimally Changed, Where the Outcome is Different…

David Kellogg Lewis (1941-2001)

Lewis (1973) Counterfactuals, CUP.

"*Minimal Changes*" & "Closeness" seem *slippery...*



- What's a minimal change?
  ~ small changes == large effects (fine margins)
  ~ large changes == small effects (cancelling factors)
- What's close will depend on sim-metric & representations
- Possible worlds get complex (true in all possible worlds, accessibility spheres, similarity analyses nested, etc.)
- Different logics have been proposed to handle issues

Stanford Encyclopedia of Philosophy (Counterfactuals)

Roots of Causality

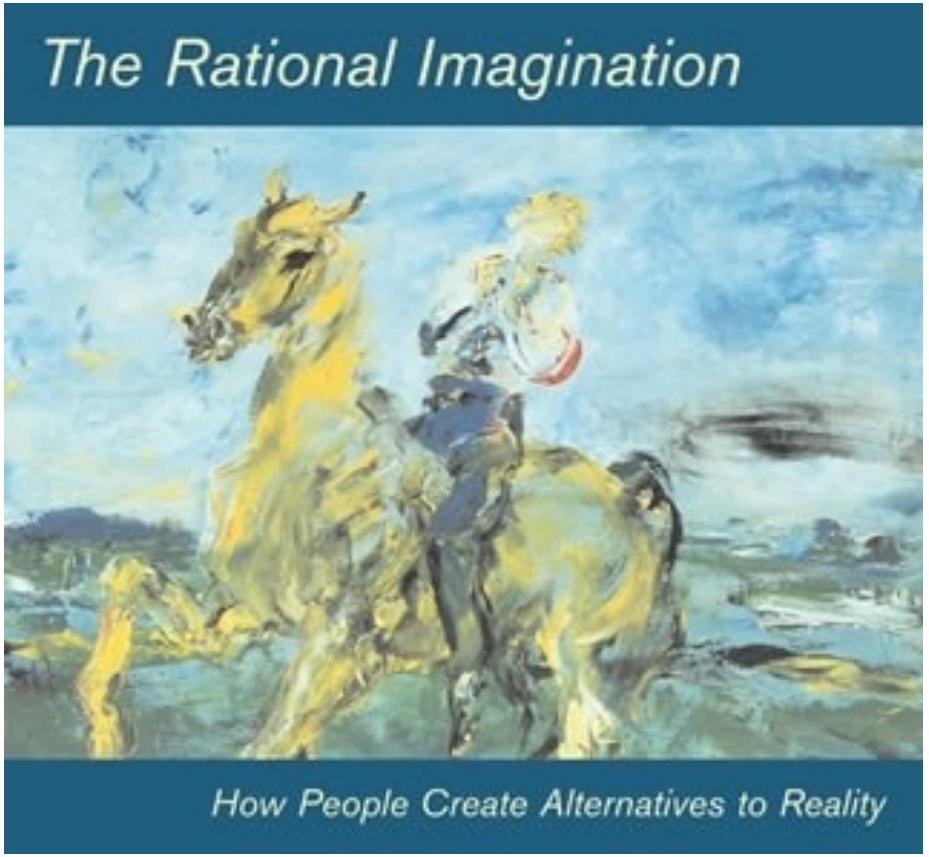
- Woodward's (2005) interventionist (in theory) account of causation in Philosophy of Science
- A causal relationship between *X* and *Y* exists if, under an ideal intervention (all other variables are controlled/held-constant), a change in *X* leads to a change in Y
- Causal claims stated using counterfactuals of the form "If *X* had not occurred, *Y* would not have occurred."
- Only way to solidly establish causality (nb. Pearl SCMs)

Woodward (2005) Making Things Happen. OUP

PSYCHOLOGY

The Fault Lines of Reality…

Daniel Kahneman (1934-2024)



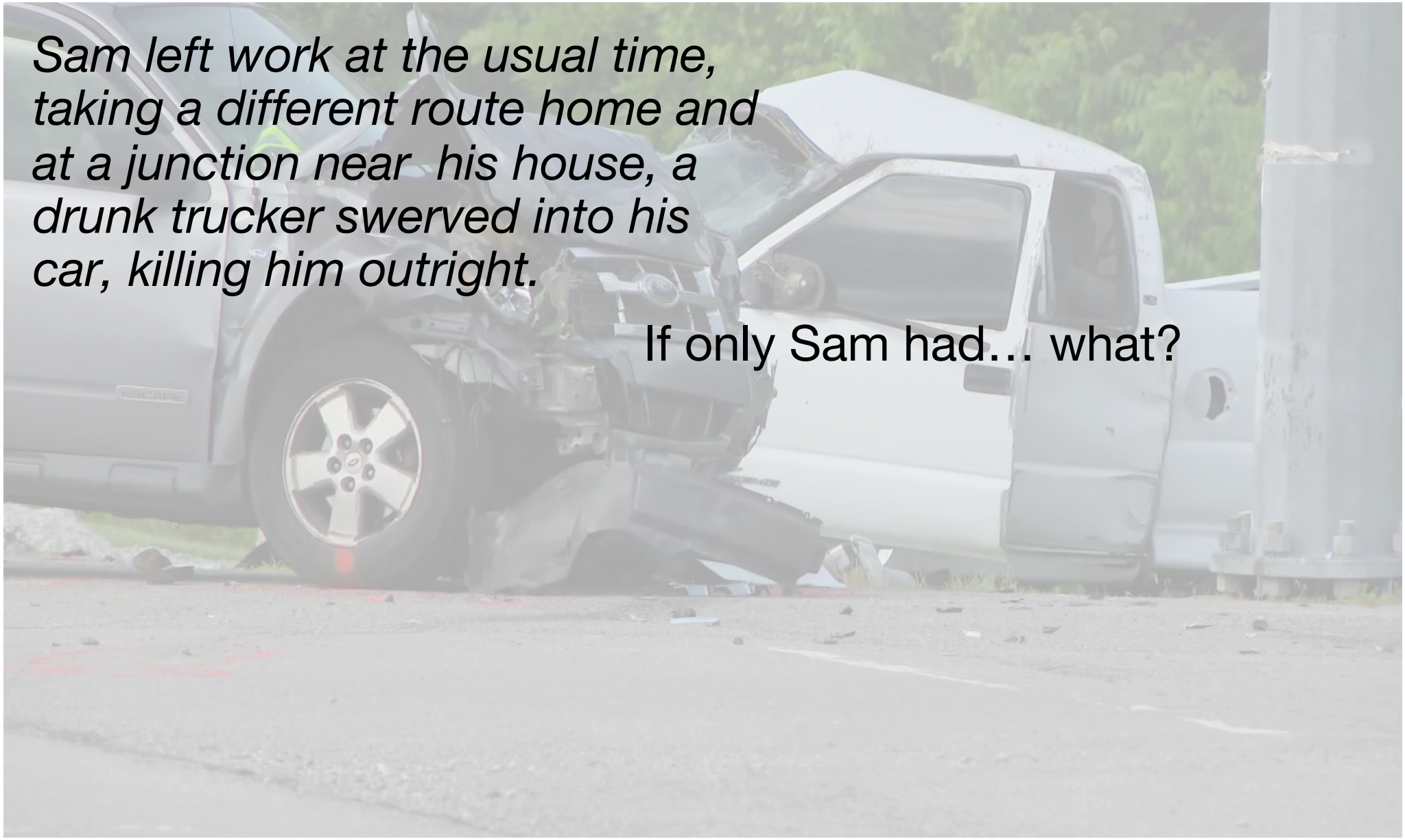
Byrne (2007) MIT Press

Counterfactual Thinking

- Critical to thinking about our past, future, causality, moral judgements (blame) with emotional import (regret, guilt, blame)
  - ~ "If only I hadn't married that idiot, I would be happy."
  - ~ "I'll do better in my next exam, if I concentrate more."
  - ~ "Heating water to 100°C caused it to boil" (implicitly invites "If the water had not been heated to 100°C, it would not have boiled")
- Thoughts about how things could have turned out differently can lead to *regret* (influencing decisions)
- *Guilt* and *self-blame* can be experienced ("if I had only visited my aging mother more, I would have made her final days better")

Byrne (2016) Annu. Rev. Psychol.

Fault Lines of Reality: What to Change?

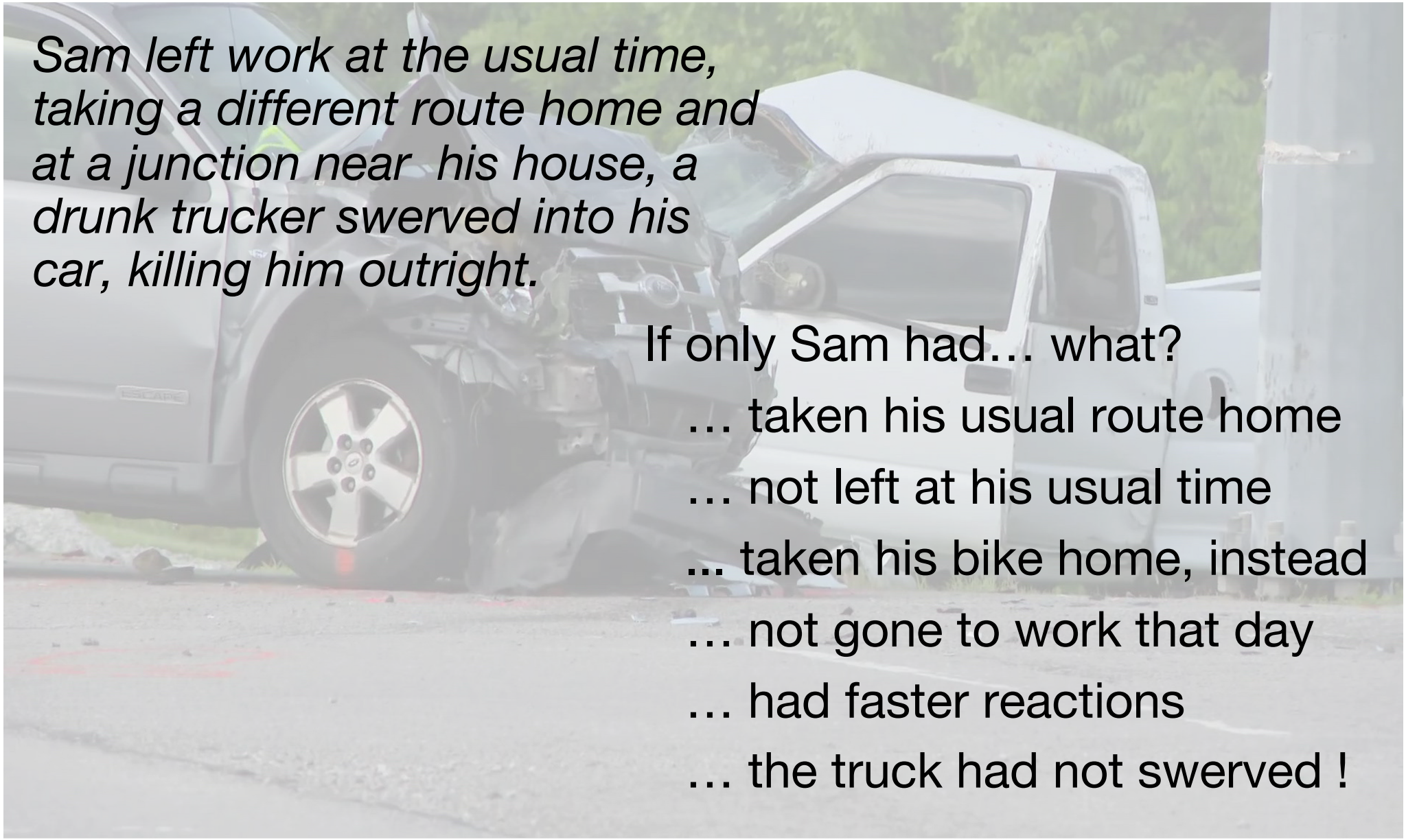


Sam left work at the usual time, taking a different route home and at a junction near his house, a drunk trucker swerved into his car, killing him outright.

If only Sam had… what?

Kahneman & Miller (1986) *Psych. Rev.*

Fault Lines of Reality: What to Change?



Sam left work at the usual time, taking a different route home and at a junction near his house, a drunk trucker swerved into his car, killing him outright.

If only Sam had… what?

… taken his usual route home

… not left at his usual time

… taken his bike home, instead

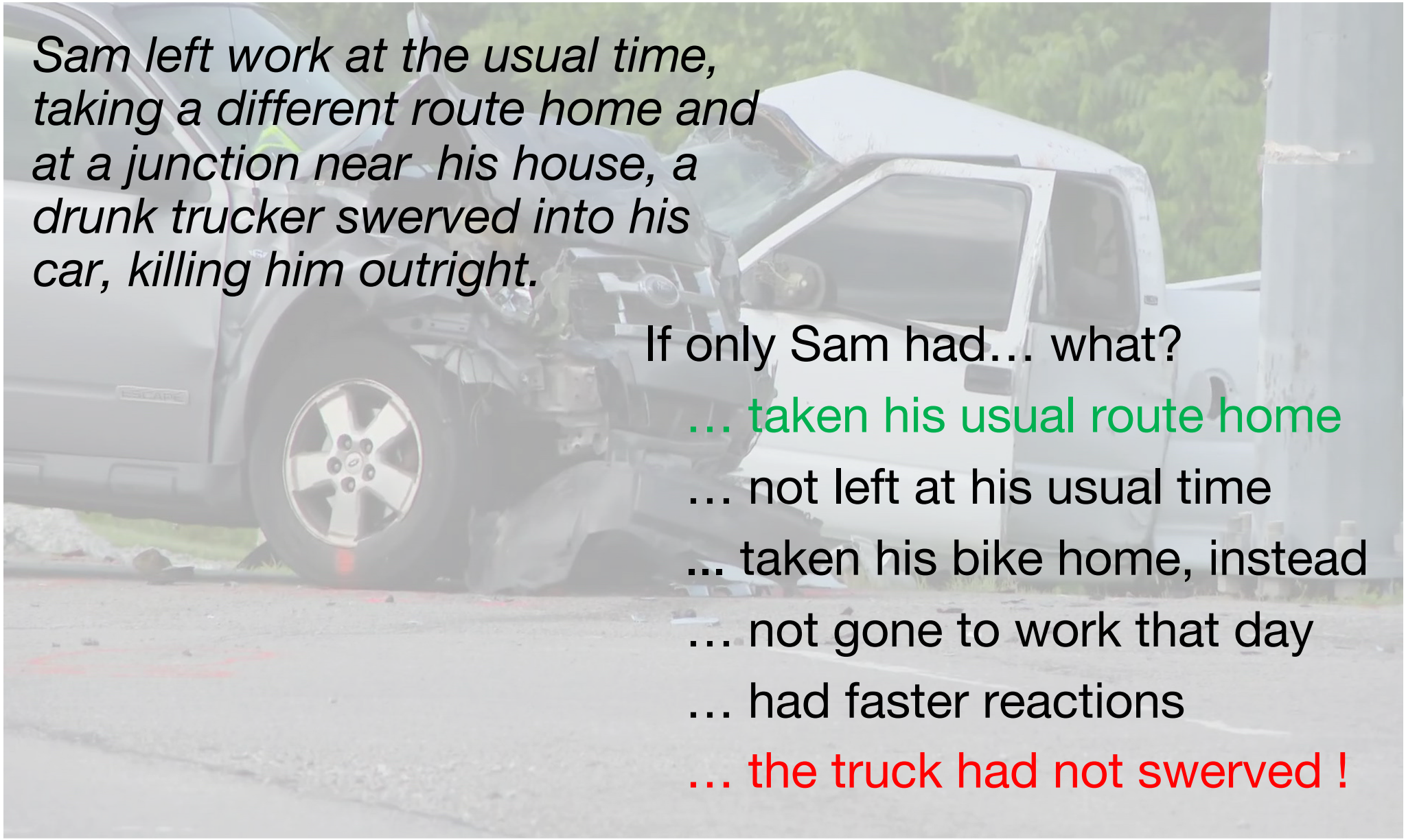
… not gone to work that day

… had faster reactions

… the truck had not swerved !

Kahneman & Miller (1986) *Psych. Rev.*

Fault Lines of Reality: What to Change?



Sam left work at the usual time, taking a different route home and at a junction near his house, a drunk trucker swerved into his car, killing him outright.

If only Sam had… what?

… taken his usual route home

… not left at his usual time

… taken his bike home, instead

… not gone to work that day

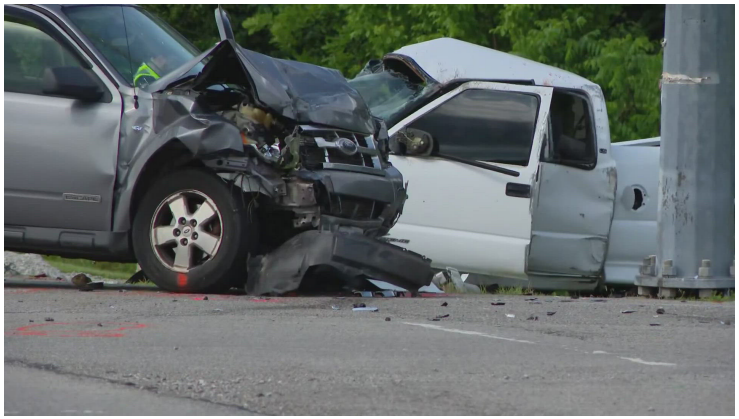
… had faster reactions

… the truck had not swerved !

Kahneman & Miller (1986) *Psych. Rev.*

Fault Lines of Reality

- People undo the same very specific things, the stress points where reality is *slippable* (Hofstadfter)
- People blame the strong cause (drunk swerving) but selected enabling causes (route taken) to change !
- They tend to change (the route):
  - ~ exceptional/non-normative events
  - ~ controllable constraints
  - ~ temporally recent to outcome
  - ~ socially unacceptable
  - ~ focal actor (victim blaming)



Byrne (2016) Annu. Rev. Psychol.

Other Effects…

- *Actions/Inactions Differ*: eg regrets about selling v holding shares
- *Causals*: people may implicitly generate CFs to establish causes of dynamic collision events
- *Directionality*: considering outcomes that could be better (upward CFs) or worse (downward CFs), negative outcomes elicit more CF thoughts
- Imagining CFs v imagining hypothetical alternatives

Epstude & Roese (2008) Pers. Soc. Psych. Rev.

Gerstenberg et al. (2021) Psych. Rev.

Psychological Theories

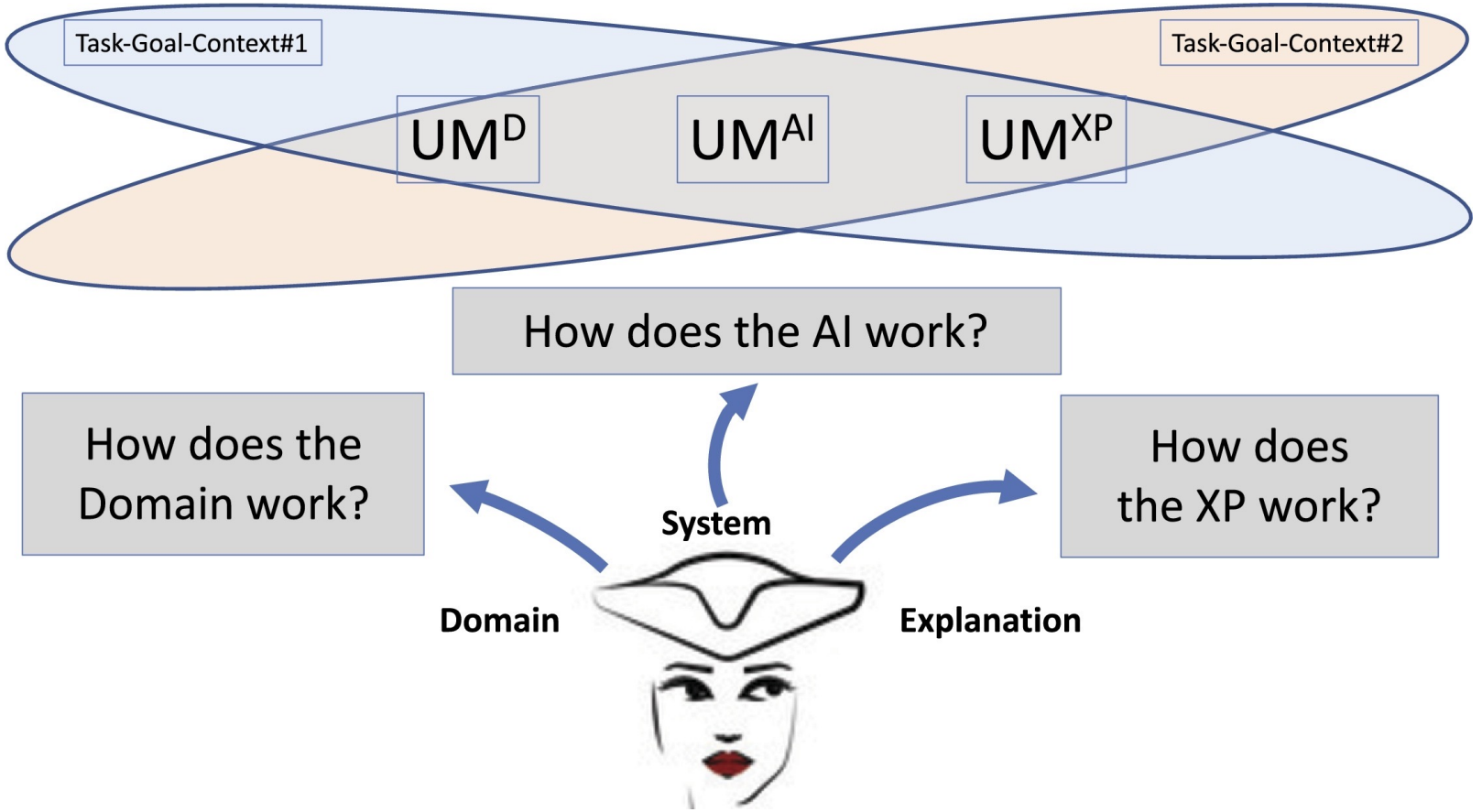
- ***Norm Theory***: Norms will tend to be immutable/given.
- ***Mental Models***: Two models are represented, what occurred and the counterfactual alternative; changes reasoning predictions and cognitive effort required
- ***Simulation***: *Running* a physical event in the counterfactual world to see contrasts with what happens
- ***Functional***: More high-level about the consequences for decision making, reasoning, preparation and planning

Byrne (2016) Annu. Rev. Psychol.

- They provide “natural” cognitive means to explain outcomes (eg. more so than say LIME)
- CFs are implicated in people’s understanding of causality, so may promote domain knowledge
- They have motivational and emotional impacts
- Lots known about them from psych research
- XAI Task (sort of) parallels the Normal-Use Task

- XAI Task is *not-identical-to* Normal-Use Task
- Features used/mutated may not be those people typically choose or, indeed, know about
- User study effects are mixed; showing poor insight into what is really impacting CF use
- Extraneous variables may sometimes be more important than XP strategy; optimal use conditions?
- Ethical issues: Every cognitive impact is exploitable!

XAI Mental Models Are Complex !



Kenny et al. (2021) Artif. Intell.

Plan for the Day

CF Tutorial

TIME	Topics
9:00 AM	***Introduction***
	Hello and Introducing Ourselves!
	Hands-on: *Trying Our Study (follow link)*
9:30 AM	***Historical Fundamentals of Counterfactuals***
	From Philosophy to XAI (via Psychology)
	Two Sample User Studies and Q&A
10:30 AM	COFFEE (10:30-11:00)
11:00 AM	***Fundamentals of Counterfactuals in AI***
	Formalisation
	Modelling Approaches & Key Constraints
11:30 AM	***Using Counterfactual Algorithms***
	Hands-on: *A Counterfactual Toolbox (AA)*
	Hands-on: *Checking Out Notebooks and Q&A*
12:00 PM	***Fundamentals of User Studies***
	User Studies I: A Simple Two-Group Design
12:30 PM	LUNCH (12:30-14:00)
2:00 PM	***Algorithmic Growth Points***
	Computational Future Directions and Q&A
2:30 PM	***More Fundamentals of User Studies***
	User Studies I: A Simple Two-Group Design (cont.)
3:00 PM	COFFEE (15:00-15:30)
3:30 PM	***From Fundamentals to an Actual User Study***
	User Studies II: A More Complex Design
	User Studies III: Even More Complex Designs
	Hands-on: *Looking At Our Study*
5:00 PM	***Closing Session, Discussion and Final Q&A***
	TUTORIAL END